3.2.2 Grouping of Clusters and Outliers

Now for each outlier O we have to find its nearest cluster C. According to the diversity measurement D1 (C, O) which is (defined in part 2.2) the distance between the cluster C and outlier O is calculated. And according to the information of the outliers in O and the information of the clusters in C the quality Q (O) (which is defined in the part 2.3) is calculated. The worst qualities of outliers are put into the set O'. Thus, for each cluster C C. According to not only the information of the clusters in C but also the information of outliers in O the quality Q (C) (defined in the part 2.3) is calculated. The worst qualities of clusters are put into set C'.

3.2.3Findboundarydatapoints

We need those data points into the clusters that are not only more distant from the centroid of the clusters but also contain the smallest number of neighboring data points as the data points of the clusters. The latest circumstances assure that this method not only promote clusters of standard geometries such as hyper spherical.

3.2.4Exchangeofdatapoints

In this step it is use to exchange the outliers and the boundary data points characteristics. Now for each outlier O in we add it into its closest cluster. For each boundary data point bp in BP we change it into a new outlier. Thus, we do not exchange the data points of the boundary between clusters are that whole data division quality will be degraded if it is carried.

Algorithm (K: Number of clusters)
Start

1. First stage
Repeat
A and B are the two equivalent constant to K, where A>B;
RS1 = A. K;
RS2 = B. K;
T1 = it is the random set having the size of RS1;
T2 = Finding K medoids from T1 having size of RS2;
H Dispatch the data points (T2);
C and O it is the cluster or the outlier ();
Until ≥ K

2. Second stage
C Merging of the Cluster (C);
Repeat
For each outlier o O do
Start
Find the nearest cluster C
Stop

Group the current set of clusters and current set of outliers in ascending order according to their qualities; Now exchange the cluster and the outlier (); O' it is the set of outliers according to the worst qualities; BP it is the set of boundary data points according to the worst qualities;
U = O' BP;
Until (U gives a constant value or the iteration ≥ Ω)
Stop.

3.3 Analysis of Time and Space

Let us assume size of data set is n. Now during the whole process we need to keep track of information on all points which collectively holds O (n) space. For the second stage or we can say the iteration stage we need space for the information of current set of clusters C and the current set of outliers O, the boundary data points for each cluster, the worst qualities of outliers and clusters in each iteration. The total amount of space needed is O (n). The time required for each iteration is

$$O\ (n + |C| log|C| + |O| log|O|)$$

particularly for the computation of the various types of qualities and sorting process. C and O. So the total amount of time required for the algorithm is O $(\Omega * (n + |C| log|C| + |O| log|O|))$ in which Ω is known as the threshold for the number of iterations.

3.4 Workflow of the Algorithm

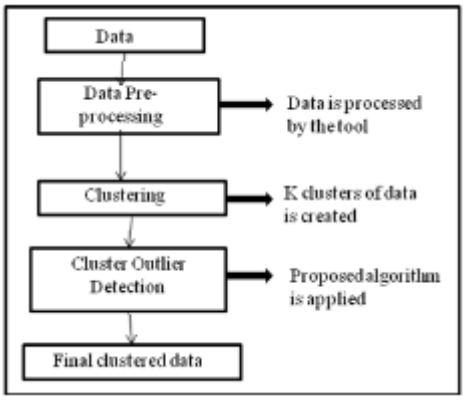

Figure1: Work flow of the algorithm

Above is the figure depicting the workflow chart of our proposed algorithm? Firstly the data set is taken and feed into the tool for pre-processing of the data and after the completion of this step K clusters of Data is being created. Then our proposed algorithm is applied on the data for cluster outlier detection and in the last we get the final clustered data.

4. Experiments and Results

In this section we have done the experiment on Ecoli data set which is taken from [11]. We conducted our experiment on Cluster 3.0 tool [12]. We use an Ecoli data set which is used to find Localization site of proteins. The dataset contains a total of 336 instances (objects) each having attributes (1 name and 7 input features). It contains eight clusters having sizes of (143, 77, 52, 35, 20, 5, 2 and 2). Now we perform our algorithm on the Ecoli data and after the experiment we observe that our proposed algorithm has the most information about the first three largest clusters.

Table 1: Clustering result of the algorithm for Ecoli data

	K=0	K=1	K=2	K=3	K=4	K=5	K=6	K=7
C (R)	143	77	52	35	20	5	2	2
C (D)	138	80	73	N	N	4	N	N
C(R)∩C(D)	132	54	50	N	N	4	N	N
Accuracy%	95.6	67.50	68.49	N	N	100	N	N
Recall Value %	92.3	70.12	96.15	N	N	80	N	N

Now from the result table we see that the data set contains 8 clusters C^* (R) for K = (0 to 7). There are three clusters which are too small so we set the clusters parameter number K to 5. And the accuracy of the detected cluster measured according to the accuracy % and the recall value %. Now for C^* (D) detected cluster and C^*(R) for the real cluster we calculate the accuracy of C^* (D) with respect to C^* (R) as $\frac{C^*(D) \cap C^*(R)}{C^*(D)}$ and recal value is $\frac{C^*(D) \cap C^*(R)}{C^*(R)}$. Hence C^* (D) is called as comparable cluster of C^*(R) if the accuracy and the recall value of C^*(D) and C^*(R) are high.

5. Conclusions

In this paper we interpolate a peculiar approach for cluster outlier detection in high dimensional data. This approach can be able to ameliorate the clusters and outliers qualities for those high dimensional data which contains noise. The clusters are ascertained and managed according to the intra-relationship within the clusters and inter-relationship between the clusters and the outliers. The whole management and modification of the clusters and outliers are done repeatedly just before a certain termination is reached. Now further dealing with the clusters and outliers as a concept of the same significance in data analysis. We also concern about the flushing the difficulties of the deficiency of match between the ground truth of the real data and their obtainable characteristics.

References

[1] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I, Statistics. 1967.
[2] D. Yu, G. Sheikholeslami and A. Zhang. Findout: Finding outliers in very large datasets. The Knowledge and Information System, (4), October 2000.
[3] Charu C. Aggarwal et al. Outlier Detection for high dimensional data. In SIGMOID Conference, 2001
[4] M.F. Jiang et al. Two phase clustering process for outlier detection. Pattern Recognition Letters 22 pages 691-700. 2001
[5] Angiuli et al. Fast outlier detection in high dimensional spaces. In the proceedings of KDD. 2002.
[6] Chi-Farn Chen, Jyh-Ming Lee. The Validity Measurement of Fuzzy C-Means Classifier for Remotely Sensed Images. In Proc. ACRS 2001 - 22nd Asian Conference on Remote Sensing, 2001.
[7] Maria Halkidi et al. A data set oriented Approach for clustering Algorithm Selection. In PKDD, 2001.
[8] Dantong Yu and Aidong Zhang. ClusterTree: Integration of Cluster Representation and Nearest Neighbor Search for Large Datasets with High Dimensionality. IEEE Transactions on Knowledge and Data Engineering (TKDE), 14(3), May/June 2003.
[9] T. Gonzalez. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38:311-322, 1985.
[10] Charu C. Aggarwal et al. Fast algorithms for projected clustering. In the proceedings of the ACM SIGMOID Conference on management of Data, pages 61-72, 1999
[11] The UCI KDD Archive [http://Kdd.ics.uci.edu]. University of California, Irvine, Department of Information and Computer Science.
[12] Michiel de Hoon. [http://bonsai.ims.u-tokyo.ac.jp], Open source clustering software. Institute of medical Science University of Tokyo.

Author Profile

Abhimanyu Kumar was born in India in 1991. He completed his B. Tech degree with distinction in Computer Science from Suresh Gyan Vihar University, Jaipur, India and now working in **Xerox Corporation** as a Business analyst in Big Data Management Team. His research interest includes A Peculiar approach for cluster outlier detection in high dimensional data.